

**AMERICAN COMMUNITY SURVEY
2012-2016 ACS 5-YEAR PUMS FILES**

ReadMe

Prepared by
American Community Survey Office
U.S. Census Bureau
January 18, 2018

I.) Overview of the Public Use Microdata Sample (PUMS)

The Public Use Microdata Sample (PUMS) contains a sample of actual responses to the American Community Survey (ACS). The PUMS dataset includes variables for nearly every question on the survey, as well as many new variables that were derived after the fact from multiple survey responses (such as poverty status). Each record in the file represents a single person, or--in the household-level dataset--a single housing unit. In the person-level file, individuals are organized into households, making possible the study of people within the contexts of their families and other household members. PUMS files for an individual year, such as 2016, contain data on approximately one percent of the United States population. PUMS files covering a five-year period, such as 2012-2016, contain data on approximately five percent of the United States population.

The PUMS files are much more flexible than the aggregate data available on American FactFinder, though the PUMS also tend to be more complicated to use. Working with PUMS data generally involves downloading large datasets onto a local computer and analyzing the data using statistical software such as R, SPSS, Stata, or SAS.

Since all ACS responses are strictly confidential, many variables in the PUMS files have been modified in order to protect the confidentiality of survey respondents. For instance, particularly high incomes are "top-coded," uncommon birthplace or ancestry responses are grouped into broader categories, and the PUMS files provide a very limited set of geographic variables (explained more below).

II.) Public Use Microdata Areas (PUMAs)

While PUMS files contain cases from nearly every town and county in the country, towns and counties (and other low-level geography) are not identified by any variables in the PUMS datasets. The most detailed unit of geography contained in the PUMS files is the Public Use Microdata Area (PUMA).

PUMAs are special non-overlapping areas that partition each state into contiguous geographic units containing no fewer than 100,000 people each. The 2012-2016 ACS 5-year PUMS files rely on PUMA boundaries that were drawn by state governments after the 2010 Census. To further investigate the use of the PUMA codes and boundaries, reference the following resources.

TIGERweb: https://tigerweb.geo.census.gov/tigerwebmain/tigerweb_main.html

- TIGERweb is an interactive mapping application published by the Census Bureau that can be used to view 2010 PUMA boundaries. To access the maps:
- Click on "TIGERweb Applications" on the upper left.
- Click "TIGERweb" on the left column.
- On the upper left, you should see choices: "Layers", "Legend", or "Task Results." Select "Layers" if it is not already selected.
- Under "Select Vintage:" choose "ACS 2016."
- Select "PUMAs, UGAs, and ZCTAs" on the left.

- Expand the “PUMAs, UGAs, and ZCTAs” box to see the choices: “2010 Census Public use Microdata Areas” and “2010 Census ZIP Code Tabulation Areas.”
- Click on the map to zoom, or move the zoom scale bar to zoom in closer to the map until the geographic outlines appear.
- Deselect “2010 Census ZIP Code Tabulation Areas” to remove them from the map.

Static Maps: <https://www.census.gov/geo/maps-data/maps/reference.html>

The Census Bureau also publishes static maps of PUMA boundaries. To access the static maps:

- Click on “Public User Microdata Areas (PUMAs)” in the middle of the screen.
- The section will expand to show three options. Choose the first option labelled “2010 Census Public Use Microdata Area (PUMA) Reference Maps”.
- The section will expand again. Choose the state you are interested in from the drop down menu and click on “Go”.
- A new webpage with the list of static maps for the relevant state will appear. Choose the geographic area of interest to see the relevant static PUMA map.

III.) PUMS Documentation

The PUMS Technical Documentation can be found by going to: <https://census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>. This page includes the following documents:

- **Subjects in the PUMS:** A list of topics included in each of the housing and population record files.
- **PUMS Data Dictionary:** Information on the variables available in the person and housing datasets. The Data Dictionary also includes the values and definitions for each variable.
- **PUMS Code Lists:** Detailed codes for variables that contain a large number of coded responses, such as ancestry and occupation.
- **PUMS Top Coded and Bottom Coded Values:** .A table by state of the top-coded values and the threshold value above which variables are top-coded and the bottom-coded and threshold below which variables are bottom-coded for relevant PUMS person and housing variables.
- **Accuracy of the PUMS:** Detailed descriptions of the sampling methodology, weighting methodology, details on protection of respondents confidentiality, and examples on how to calculate standard errors using the design factor methodology are provided in this document.
- **PUMS Estimates for User Verification:** PUMS estimates for selected housing and population characteristics. Data users may calculate these characteristics independently

and then verify their process is correct by matching to these estimates. The standard errors and margins of error, calculated using replicate estimates, are also provided in this document.

IV.) Getting PUMS data

The PUMS datasets are available at:

ACS Website:

PUMS files can be accessed via the ACS website: <https://www.census.gov/programs-surveys/acs/data/pums.html>.

American FactFinder:

PUMS Files are also accessible via American FactFinder: <https://factfinder.census.gov>.

DataFerrett:

PUMS data is also available via the Census Bureau's DataFerrett, which has the additional feature of being able to make tables and perform basic analysis online. This tool is particularly useful for researchers who need a quick statistic or do not have access to statistical software. DataFerrett is available at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/dataferrett.html>.

Note: DataFerrett only works with Microsoft Internet Explorer and the ESR version of Mozilla Firefox. DataFerrett does not work with Microsoft Edge or Google Chrome.

V.) PUMS Files Structure

The ACS questionnaire contains "household" items that are the same for all members of the household (such as the number of rooms in the home) and "person" items that are unique for each household member (such as age, sex, and race). The ACS PUMS files are made available in this same structure. Researchers who are analyzing only household-level items can use the household files, whereas those using only person-level variables can use the person-level files.

Data users should note that PUMS files containing data for the entire United States (in contrast to individual state and state-equivalent files) are separated into multiple data files. These files are separated by unique records that must be concatenated in order to create a complete file for the nation. For example, users downloading the 2012-2016 ACS 5-year PUMS person files for the nation will notice an "a" file, "b" file, "c" file and "d" file. Each file contains about one-fourth of the population records in the 2012-2016 ACS 5-year PUMS dataset of the United States. Below are instructions for concatenating the four PUMS person-level files, along with SAS pseudo-code in italics.

Concatenate the four **person-level** files. In SAS, you can use the set statement:

```
data population;
set psam_pusa psam_pusb psam_pusc psam_pusd;
run;
```

The 2012-2016 ACS 5-year PUMS files of the United States Housing Records also contain an “a” file, “b” file, “c” file and “d” file. To create a complete housing-level file, the four files must be concatenated. Below are instructions for concatenating the four PUMS household-level files, along with a SAS pseudo-code in italics

Concatenate the two **household-level** file. In SAS, you can use the set statement:

```
data housing;
set psam_husa psam_husb psam_husc psam_husd;
run;
```

Some data users will need to use household and person items together. For instance, in order to analyze how the number of rooms in a home varies by a person’s age, the merging of the household and person files will be required. This merger must rely on the SERIALNO variable, which is the same in the household and person files. Below are instructions for merging the housing and population PUMS files, along with a SAS pseudo-code in italics

Use the variable SERIALNO to merge population and housing files.

First, make sure the files are sorted by SERIALNO.

```
proc sort data=population;
by serialno;
run;
```

```
proc sort data=housing;
by serialno;
run;
```

Next, merge the two files together using SERIALNO as a merge key. Note that the housing files contain data for vacant housing units.

```
data combined;
merge population (in=pop) housing;
by serialno;
if pop; * retains all person data, and omits
vacant housing units *;
run;
/*In SAS, the 'in=' option will allow you to
keep only those housing units that have
people*/
/*This SAS statement keeps only those housing
units that were in the population file*/
```

You should not merge the files unless the estimates you want require a merge. Note that there are many estimates that may be tabulated from the person file and from the household file without any

merging. The suggested merge will create a person level file, so that the estimate of persons may be tallied within categories from the household file and the person weights should be used for such tallies.

Note: Housing characteristics cannot be tallied from this merged file without extra steps to ensure that each housing weight is counted only once per household.

VI.) PUMS Weights

The ACS PUMS is a weighted sample, and weighting variables must be used to generate accurate estimates and standard errors. The PUMS files include both population weights and household weights. Population weights should be used to generate statistics about individuals, and household weights should be used to generate statistics about housing units. The weighting variables are described briefly below.

PWGTP: Person's weight for generating statistics on individuals (such as age).

WGTP: Household weight for generating statistics on housing units and households (such as average household income).

WGTP1-WGTP80 and PWGTP1-PWGTP80: Replicate weighting variables, used for generating the variance, standard errors, and margins of errors for households or individuals using the successive difference replication (SDR) method.

PWGTP and WGTP may be used both to generate the point estimates and to generate standard errors when using a generalized variance formula (GVF). Replicate weights may only be used just to calculate "direct standard errors" using the SDR method. Direct standard errors are expected to be more accurate than GVF standard errors, although they may require more computer resources to calculate. Generalized standard errors are explained in more detail in the PUMS Accuracy of the Data document, which can be found at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

The formula for the SDR method is also in the PUMS Accuracy of the Data document. Data users may also wish to consult the Technical Documentation for the Variance Replicate Tables for practical examples and more information on using the SDR methodology. The Variance Replicate Table documentation pertains to ACS data, but can be applied to PUMS data as well. This documentation is available at: <https://www.census.gov/programs-surveys/acs/technical-documentation/variance-tables.html>.

Each housing unit and person record contains 80 replicate weights. To use the replicate weights to calculate an estimate of the direct standard error, first form the estimate using the full PUMS weight. Then form the estimate using each of the 80 replicate weights--providing both the full PUMS estimate and 80 replicate estimates. These estimates are then entered into the following formula, which is explained in more detail in the "Accuracy of the PUMS" document:

$$SE(X) = \sqrt{\frac{4}{80} \sum_{r=1}^{80} (X_r - X)^2}$$

Where X_r is a replicate estimate from X_1 to X_{80} , and X is the full PUMS weighted estimate.

The technical explanation of the ACS replicate weights may also be found in Chapter 12 of the Design and Methodology document found at: <https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html>.¹

Please note that many estimates generated with PUMS may differ from estimates for the same characteristics published on American FactFinder. These differences are due to the fact that the PUMS data is a subset of the full ACS microdata (with confidentiality protections applied to it), as well as additional PUMS edits. More information on the PUMS sample design is available in the "PUMS Accuracy of the Data" document, which can be found at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

VII.) Variable Changes in the 2012-2016 ACS 5-year PUMS files

The 2012-2016 ACS 5-year PUMS includes the variables that were in the 1-year PUMS from 2012 through 2016.² See the 2012-2016 5-year PUMS Data Dictionary for a comprehensive listing of the 2012-2016 ACS 5-year PUMS variables and their values. The data dictionary is available at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

The following lists and table provides some information about relevant changes to the variables in the 2012-2016 ACS 5-year PUMS, as compared to the 2011-2015 ACS 5-year PUMS. For more details on the changes, see both the 2011-2015 and 2012-2016 ACS 5-year PUMS Data Dictionaries and Code Lists from the PUMS Technical Documentation page: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

Variables added since previous 5-year PUMS: RACNH, RACPI.

Variables deleted since previous 5-year PUMS: ANC1P05, ANC2P05, CITWP05, LANP05, MARHYP05, MIGPUMA00, MIGSP05, OCCP10, POB05, POWPUMA00, POWSP05, PUMA00, RAC2P05, RAC3P05, RACNHPI, SOCP10, YOEP05.

Variables with new or modified codes since the previous 5-year PUMS: ADJHSG, ADJINC, BUS, CITWP, CONP, LANP, MARHYP, PLM, SERIALNO, SVAL, TOIL, YBL, YOEP, FBUSP, FTOILP.

¹ For more information on the theoretical basis, please reference: Fay, R. and Train, G. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Section on Government Statistics, American Statistical Association, pp. 154-159, 1995.

² The new household language and computer and internet related variables are not available on the 2012-2016 5-year PUMS, as they were only included in the 2013, 2014, 2015, and 2016 ACS 1-year PUMS.

Variables with cosmetic changes to variable labels or value labels: ANC1P, ANC2P, FER, LANP, MIGPUMA, MIGSP, OCCP, POBP, POWPUMA, POWSP, PUMA, RAC2P, RAC3P, SOCP, SRNT, TEL, FTELP.

Detailed Changes by Variable (as compared to 2011-2015 ACS 5-year PUMS Data Dictionary)

Variable	Description of Change
<i>New Variables</i>	
RACNH	New variable added because 5 years of data now available.
RACPI	New variable added because 5 years of data now available.
<i>Variables with Code Changes</i>	
BUS	BUS is not available for year 2016. ³ Cases from 2016 are assigned a value of '9'. Users should not calculate a 5-year estimate from the 2012-2016 PUMS Files, as the lack of 2016 data will give inaccurate results.
FBUSP	Variable set to zero ("No") for year 2016.
SVAL	Description updated in data dictionary, as business on property is no longer a component in SVAL as of 2016. ³ The values of this variable for years 2012, 2012, 2014 and 2015 still require business on property. See User Note for more information. ⁴
TOIL	TOIL is not available for year 2016. ⁵ Cases from 2016 are assigned a value of '9'. Users should not calculate a 5-year estimate from the 2012-2016 PUMS Files, as the lack of 2016 data will give inaccurate results.
FTOILP	Variable set to zero ("No") for year 2016.
PLM	Description updated in data dictionary, as complete plumbing no longer requires a flush toilet as of 2016. ⁵ The values of this variable for years 2012, 2013, 2014 and 2015 still require a flush toilet. See User Note for more information. ⁴
CONP	Description updated in data dictionary, as the universe for CONP now excludes all vacant housing units. As such, all vacant units have CONP set to blank. See Errata Note 117 for more information. ⁴
LANP	Language codes updated and expanded to 4-digits. See Data Dictionary for new definitions.
TEL	Data for this variable suppressed for selected PUMAs. Variable is set to '8' for suppressed PUMAs. See Data Dictionary for an updated note.
<i>Variables with Annual Code Updates</i>	
ADJHSG	Adjustment factors updated.
ADJINC	Adjustment factors updated.
SERIALNO	Range updated for data year 2016.
YBL	Most current year (2016) added.
<i>Variables with Annual Code Updates and Name Changes</i> ⁶	
CITWP	Most current year (2016) added. Formerly named CITWP12.
MARHYP	Most current year (2016) added. Formerly named MARHYP12.

³ The business on property question was removed from the survey for 2016. Variables FBUSP and SVAL also affected.

⁴ User Notes and Errata can be found at: <https://www.census.gov/programs-surveys/acs/technical-documentation.html>.

⁵ The flush toilet question was removed from the survey for 2016. Variables FTOIL and PLM also affected.

⁶ For more information on variable name changes, see the section "Removal and Renaming of Dual Vintage Variables."

Variable	Description of Change
<i>Variables with Annual Code Updates and Name Changes</i> ⁷	
YOEP	Most current year (2016) added. Formerly named YOEP12.
<i>Variables with Description Changes Only</i>	
SRNT	Definition clarified in Data Dictionary.
FER	Text of note updated in Data Dictionary.
FTELP	Definition in Data Dictionary updated to reflect new telephone question wording.
<i>Variable with Name Changes Only</i> ⁷	
ANC1P	Formerly named ANC1P12.
ANC2P	Formerly named ANC2P12.
MIGPUMA	Formerly named MIGPUMA10.
MIGSP	Formerly named MIGSP12.
OCCP	Formerly named OCCP12.
POBP	Formerly named POBP12.
POWPUMA	Formerly named POWPUMA10.
POWSP	Formerly named POWSP12.
PUMA	Formerly named PUMA10.
RAC2P	Formerly named RAC2P12.
RAC3P	Formerly named RAC3P12.
SOCP	Formerly named SOCP12.

Removal and Renaming of Dual Vintage Variables

In data year 2012, a set of dual vintage variables were introduced to the PUMS 5-year file.⁸ The 2012-2016 ACS 5-year PUMS dataset is the first 5-year product to contain five years of data with the 2012 vintage variables. On the 2012-2016 files, these variables have been renamed, with the year now removed from the variable name. For instance, variable POBP12 on the 2011-2015 PUMS 5-year dataset is named POBP on the 2012-2016 dataset.

NAICSP and INDP are not dual vintage variables, as the 2012 categories were crosswalked to the new values used for the 2013, 2014, 2015, and 2016 1-year PUMS. Likewise, LANP is not listed as a dual vintage variable, though it changed from a three-digit code to a four-digit code in 2016. LANP codes from 2012, 2013, 2014, and 2015 were crosswalked to the four-digit codes for the 2012-2016 ACS 5-year PUMS files. For more information on variable crosswalks, see the PUMS Code Lists document at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html>.

⁷ For more information on variable name changes, see the section “Removal and Renaming of Dual Vintage Variables.”

⁸ Former datasets contained dual vintage variables, where one variable vintage was used for data from 2011 and earlier and a second vintage for 2012 or later. Each vintage of the variable contained the same root name, with a year reference suffix to distinguish them from each other (i.e. POBP05 and POBP12). For more information on dual vintage variables, see the 2011-2015 ACS 5-year PUMS ReadMe document at: <https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.2015.html>.

Variables for Puerto Rico

There is a set of dual vintage variables that applies to Puerto Rico only: RWAT and RWATPR. Beginning with the 2013 data year, the hot and cold running water variable RWAT is set to '9' for not applicable in Puerto Rico, while a new variable RWATPR (running water) replaces it. For years previous to 2013, use RWAT.

Variables with Suppressed Values

FER: Problems in the collection of data on women who gave birth in the past year (FER) led to suppressing this variable in 59 PUMAs within states Florida, Georgia, Kansas, Montana, North Carolina, Ohio and Texas for data year 2012. A code of 8 was applied to these cases for data year 2012 only.

TEL: Problems in the collection of data on the availability of telephone service (TEL) led to suppressing this variable in various PUMAs in 2012, 2015 and 2016. In each year, a code of '8' was applied to affected cases for the affected data year only.

For data year 2012, the variable TEL was suppressed in six PUMAs in Georgia.

For data year 2015, the variable TEL was suppressed in five PUMAs: One PUMA in Arkansas, two in Florida, one in Kentucky and one in Wisconsin.

For data year 2016, the variable TEL was suppressed in fourteen PUMAs: One PUMA in Illinois, three in North Carolina, two in South Carolina and eight in Texas.

Variables Discontinued in 2016

Variables BUS, FBUSP, TOIL and, FTOILP were discontinued in 2016 due to questionnaire changes. However, these variables remain on the 2012-2016 file. These variables retain their values for data from years 2012, 2013, 2014 and 2015. Variables BUS and TOIL were given a value of '9' for data from year 2016 on the 2012-2016 PUMS files; variables FBUS and FTOIL were assigned a value of zero. As the 2012-2016 file does not provide values for all five data years, these variables should not be used to create estimates. For more information, see the User Note at: <https://www.census.gov/programs-surveys/acs/technical-documentation.html>.

VIII.) Additional Information

The Census Bureau occasionally provides corrections or updates to PUMS files. Users are notified of these updates via:

The Census Bureau's E-mail Updates System:

https://service.govdelivery.com/accounts/USCENSUS/subscriber/new?category_id=USCENSUS_C12.

The ACS Errata Page:

<https://www.census.gov/programs-surveys/acs/technical-documentation/errata.html>.

Please contact acso.users.support@census.gov with any PUMS-related questions.